Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A modified local quadratic approximation algorithm for penalized optimization problems

Q1 Sangin Lee [a], Sunghoon Kwon [b], Yongdai Kim [c,*]

[a] University of Iowa, United States

[b] Konkuk University, Republic of Korea

[c] Seoul National University, Republic of Korea

A R T I C L E   I N F O

Article history:
Received 26 May 2014
Received in revised form 23 August 2015
Accepted 27 August 2015
Available online xxxx

Keywords:
Local quadratic approximation
ℓ_1-penalization
Nonconvex penalization
LASSO
SCAD
MCP

A B S T R A C T

In this paper, we propose an optimization algorithm called *the modified local quadratic approximation algorithm* for minimizing various ℓ_1-penalized convex loss functions. The proposed algorithm iteratively solves ℓ_1-penalized local quadratic approximations of the loss function, and then modifies the solution whenever it fails to decrease the original ℓ_1-penalized loss function. As an extension, we construct an algorithm for minimizing various nonconvex penalized convex loss functions by combining the proposed algorithm and convex concave procedure, which can be applied to most nonconvex penalty functions such as the smoothly clipped absolute deviation and minimax concave penalty functions. Numerical studies show that the algorithm is stable and fast for solving high dimensional penalized optimization problems.

1. Introduction

The ℓ_1-penalized estimations play important roles in high-dimensional data analysis. For example, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) in linear regression models, which minimizes ℓ_1-penalized sum of squares of residuals. The LASSO produces sparse solutions and achieves higher prediction accuracy than classical stepwise variable selection methods. There exist many optimization algorithms for minimizing ℓ_1-penalized sum of squares of residuals. For example, quadratic programming (QP) techniques were considered by Tibshirani (1996) and Osborne et al. (2000). Efron et al. (2004) developed the least angle regression and selection (LARS) algorithm that can find the whole solution path with respect to a tuning parameter. A coordinate descent (CD) algorithm was introduced by Friedman et al. (2007) that is simple but fast.

In addition to linear regression models, the ℓ_1-penalized approaches have been applied to various high-dimensional statistical models; wavelet analysis (Chen et al., 1999), kernel machine methods (Gunn and Kandola, 2002), smoothing splines (Zhang et al., 2004), logistic regression models (Park and Hastie, 2007) and multi-class logistic regression models (Kim et al., 2006). These models require optimization algorithms for minimizing ℓ_1-penalized convex loss functions that are not quadratic functions. Lokhorst et al. (2006) developed a modified QP technique for logistic regression models, and Kim et al. (2008b) proposed a gradient decent algorithm for general convex loss functions with an ℓ_1-constraint. Using the idea of the LARS algorithm, Park and Hastie (2007) proposed an approximated path-finding algorithm for the ℓ_1-penalized generalized linear models. Friedman et al. (2010) proposed a local quadratic approximation (LQA) algorithm for

* Corresponding author.
E-mail address: ydkim0903@gmail.com (Y. Kim).

http://dx.doi.org/10.1016/j.csda.2015.08.019
0167-9473/

ℓ_1-penalized generalized linear models that iteratively minimizes the ℓ_1-penalized local quadratic approximation of the likelihood function using the CD algorithm of Friedman et al. (2007).

In this paper, we propose an optimization algorithm called the modified local quadratic approximation (MLQA) algorithm for minimizing various ℓ_1-penalized convex loss functions. The MLQA algorithm modifies the LQA algorithm to have a descent property (i.e. the objective function always decreases). Hence, the proposed MLQA algorithm always converges even when the LQA fails to converge. Moreover, the modification step in the MLQA requires a one-dimensional convex optimization that can be easily done by a line search, and hence the additional computational burden is minimal. As an extension, we develop an algorithm called the CCCP–MLQA algorithm for minimizing various nonconvex penalized convex optimization problems by combining the MLQA algorithm and convex concave procedure (CCCP) of Yuille and Rangarajan (2003). The CCCP–MLQA algorithm can be applied to most of nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001) and minimax concave (MC) penalty of Zhang (2010).

There are various modified versions of the LQA algorithm. An example is to modify the Hessian matrix, where Krishnapuram et al. (2005) suggested to use an upper bound of the Hessian instead of the Hessian itself to approximate the objective function by a quadratic function. However, the algorithm of Krishnapuram et al. (2005) is not applicable when the upper bound of the Hessian matrix does not exist (e.g. Poisson regression). Yuan et al. (2012) proposed an improved version of the LQA algorithm of Friedman et al. (2010) by adapting a line search method at every iteration of the inner loop in the CD algorithm. The number of line searches, however, is proportional to the dimension of parameters and so not efficient for high dimensional model. Modifications of the LQA algorithm for general nonlinear optimization problems can be found in De Borst et al. (2012), Bishop (1995) and Böhning and Lindsay (1988).

The paper is organized as follows. We introduce the proposed MLQA algorithm in Section 2 with the proof of its descent property. Section 3 presents stability and efficiency of the MLQA algorithm through simulations. Section 4 introduces the CCCP–MLQA algorithm and discussions follow in Section 5.

2. MLQA algorithm for ℓ_1-penalized estimation

2.1. Review of LQA algorithm

Consider the problem of obtaining the ℓ_1-penalized estimator $\hat{\boldsymbol\beta} \in \mathbb{R}^p$:

$$\hat{\boldsymbol\beta} = \arg\min_{\boldsymbol\beta} Q_\lambda(\boldsymbol\beta), \tag{1}$$

where $Q_\lambda(\boldsymbol\beta) = L(\boldsymbol\beta) + \lambda\|\boldsymbol\beta\|_1$, $L : \mathbb{R}^p \to \mathbb{R}$ is a convex loss function, $\lambda > 0$ is a given tuning parameter and $\|\cdot\|_1$ denotes the ℓ_1-norm operator. If the loss function L is a quadratic function of $\boldsymbol\beta$, we can directly apply the LARS or CD algorithm (Efron et al., 2004; Friedman et al., 2007) to obtain $\hat{\boldsymbol\beta}$. However, it is not easy to find $\hat{\boldsymbol\beta}$ if the loss function L is not quadratic.

Algorithm 1 The LQA algorithm for minimizing $Q_\lambda(\boldsymbol\beta)$

Set an initial estimator $\tilde{\boldsymbol\beta}$
repeat
  Update $\tilde{\boldsymbol\beta}$ with $\hat{\boldsymbol\beta}^a = \arg\min_{\boldsymbol\beta} \tilde{Q}_\lambda(\boldsymbol\beta; \tilde{\boldsymbol\beta})$
until convergence

Given a current estimator $\tilde{\boldsymbol\beta}$, consider a local quadratic approximation $\tilde{L}$ of L around $\tilde{\boldsymbol\beta}$:

$$\tilde{L}(\boldsymbol\beta; \tilde{\boldsymbol\beta}) = L(\tilde{\boldsymbol\beta}) + \nabla L(\tilde{\boldsymbol\beta})^T(\boldsymbol\beta - \tilde{\boldsymbol\beta}) + (\boldsymbol\beta - \tilde{\boldsymbol\beta})^T\nabla^2 L(\tilde{\boldsymbol\beta})(\boldsymbol\beta - \tilde{\boldsymbol\beta})/2, \tag{2}$$

where $\nabla L(\boldsymbol\beta) = \partial L(\boldsymbol\beta)/\partial\boldsymbol\beta$ and $\nabla^2 L(\boldsymbol\beta) = \partial^2 L(\boldsymbol\beta)/\partial\boldsymbol\beta^2$. The LQA algorithm presented in Algorithm 1 finds $\hat{\boldsymbol\beta}$ by iteratively updating $\tilde{\boldsymbol\beta}$ with

$$\hat{\boldsymbol\beta}^a = \arg\min_{\boldsymbol\beta} \tilde{Q}_\lambda(\boldsymbol\beta; \tilde{\boldsymbol\beta}), \tag{3}$$

until convergence, where $\tilde{Q}_\lambda(\boldsymbol\beta; \tilde{\boldsymbol\beta}) = \tilde{L}(\boldsymbol\beta; \tilde{\boldsymbol\beta}) + \lambda\|\boldsymbol\beta\|_1$. If the current estimator $\tilde{\boldsymbol\beta}$ is sufficiently close to $\hat{\boldsymbol\beta}$, the quadratic approximation works well, so that the LQA algorithm has a descent property. That is $Q_\lambda(\hat{\boldsymbol\beta}^a) \le Q_\lambda(\tilde{\boldsymbol\beta})$.

The LQA algorithm is easy to be implemented for sufficient smooth convex loss functions since it requires only the first and second derivatives. Moreover, the ℓ_1-penalized quadratic function $\tilde{Q}_\lambda(\boldsymbol\beta; \tilde{\boldsymbol\beta})$ in (3) is easy to be minimized by using the LARS or CD algorithms. For the completeness, we present the CD algorithm for the LQA in Algorithm 2. In the algorithm, $\tilde\eta_j$ and $\tilde\zeta_j$ are the jth element of $\nabla L(\tilde{\boldsymbol\beta})$ and the jth diagonal entry of $\nabla^2 L(\tilde{\boldsymbol\beta})$, respectively, and $\tilde{\boldsymbol\alpha}_j^a$ and $\tilde{\boldsymbol\gamma}_j$ are the vectors obtained by deleting the jth element from the $\tilde{\boldsymbol\beta}^a$ and the jth column vector of $\nabla^2 L(\tilde{\boldsymbol\beta})$, respectively. See Friedman et al. (2010) for details.